BAYESIAN STATISTICS 7, pp. 000–000
J. M. Bernardo, M. J. Bayarri, J. O. Berger, A. P. Dawid,
D. Heckerman, A. F. M. Smith and M. West (Eds.)

Markov chain Monte Carlo-based approaches for inference in computationally intensive inverse problems

DAVE HIGDON
Los Alamos National Laboratory and Duke University
dhigdon@lanl.gov

HERBIE LEE
Duke University
herbie@stat.duke.edu

CHRIS HOLLOMAN
Duke University
chris@stat.duke.edu

SUMMARY

A typical setup for many inverse problems is that one wishes to update beliefs about a spatially dependent set of inputs x given rather indirect observations y. Here, the inputs and observed outputs are related by the complex physical relationship $y = \zeta(x) + \epsilon$. Applications include medical and geological tomography, hydrology, and the modeling of physical and biological systems. We consider applications where the physical relationship $\zeta(x)$ can be well approximated by detailed simulation code $\eta(x)$.

When the forward simulation code $\eta(x)$ is sufficiently fast, Bayesian inference can, in principle, be carried out via Markov chain Monte Carlo (MCMC). Difficulties arise for two main reasons:

- Even though the code may accurately represent the physical process, there are a large number of unknown, but required, inputs that must be calibrated to match the observed data y.
- The computational burden of the fastest available forward simulators is often large enough that approaches for speeding up the MCMC calculations are required.

This paper develops approaches for specifying effective low-dimensional representations of the inputs x along with MCMC approaches for sampling the posterior distribution. In particular we consider augmenting the basic formulation with fast, possibly coarsened, formulations to improve MCMC performance. This approach can be very easily implemented in a parallel computing environment. We give examples in single photon emission computed tomography and in hydrology.

Keywords: MULTIGRID MARKOV CHAIN MONTE CARLO, METROPOLIS COUPLED MARKOV CHAIN MONTE CARLO, SPATIAL STATISTICS, DISTRIBUTED COMPUTING.

1. INTRODUCTION

A typical setup for many inverse problems is that one wishes to update beliefs about a spatially dependent set of inputs x given indirect observations $y = (y_1, \ldots, y_n)^T$. Here the inputs and observed outputs are related by the complex physical relationship $y = \zeta(x) + \epsilon$ where $\zeta(x)$ denotes the actual physical system at the true, but unknown state $x = (x_1, \ldots, x_m)$, and ϵ denotes sampling error. Many such systems can be approximated by detailed computer simulation code $\eta(x)$. A very incomplete list of applications includes medical tomography (Weir, 1997), geological tomography (Andersen *et al.* 2001), hydrology (Lee *et al.* 2002), petroleum engineering (Hegstad and Omre 2001: Craig *et al.* 2001), as well as a host of other physical, biological, or social systems. The observed data

$$y = \zeta(x) + \epsilon$$

are modeled statistically by

$$y = \eta(x) + e$$

where the discrepancy term e accounts for both sampling error and mismatch between the simulator $\eta(x)$ and reality $\zeta(x)$:

$$e = \zeta(x) - \eta(x) + \epsilon.$$

The goal is to use the observed data y to make inference about the spatial input parameters x – in particular, to characterize the uncertainty about x.

The likelihood $L(y|x, \theta_y)$, which may depend on additional parameters held in θ_y, is then specified to account for both mismatch and sampling error. It is worth noting here that the data come only from a single experiment. So there is no opportunity to obtain data from additional experiments for which some controlable inputs have been varied. Because of this, there is little hope of modeling the mismatch term $\zeta(x)-\eta(x)$ separately from the sampling error as is often done in the statistical analysis of complex computer code outputs (Kennedy and O'Hagan, 2001). Therefore, the likelihood specification will often need to be done with some care, incorporating the modeler's judgement about the appropriate size and nature of the mismatch term.

We consider systems for which the model input paramaters x denote a spatial field or image. For example, in single photon emission computed tomography (SPECT) the image intensity x denotes blood flow within a region of the body; in a hydrologic application, x might give the spatial distribution of hydraulic conductivities or permeability. The simulator requires gridded inputs and the resolution of the grid is a pre-specified input to the simulator. The spatial prior for x, $\pi(x|\theta_x)$, will typically include an additional parameter vector θ_x to control x. The parameter θ_x may then be treated as fixed or have a prior of its own $\pi(\theta_x)$. Both modeling and computing considerations go into specification of $\pi(x|\theta_x)$, which is discussed in the following section.

The resulting posterior is then given by

$$\pi(x, \theta) \propto L(y|\eta(x), \theta_y) \times \pi(x|\theta_x) \times \pi(\theta)$$

where θ holds both nuisance parameters (θ_y, θ_x), This posterior can, in principle, be explored via Markov chain Monte Carlo (MCMC). However the combined effects of the high dimensionality of x and the computational demands of the simulator make implementation difficult in practice. By itself, the high dimensionality of x isn't necessarily a problem. MCMC with single-site updating has been carried out with relative ease

in large image applications. However, a high dimensional input vector x does make it quite difficult to build any sort of statistical model $\hat{\eta}(x)$ to approximate the simulator as in Sacks *et al.* (1989) or Kennedy and O'Hagan (2001). Any MCMC implementation using a single-site updating scheme is impractical since it will require m forward runs for a single update scan through all the parameters. In addition, a simulator may require a fine grid to ensure satisfactory numerical performance, but the numerical error may unduly affect the small changes in output y when only a single component of x has been updated. The use of higher dimensional proposals has proven somewhat successful (Oliver *et al.* 1997; Lee *et al.* 2002), especially when some direct measurements on x are available as in Hegstad and Omre (2001). A similar strategy that we have found effective is to reparameterize x; this is discussed in Section 2.

To deal with the computational burden of the forward simulator $\eta(x)$, Section 3 lays out a Metropolis coupled MCMC (Geyer, 1991) implementation that simultaneously runs chains to sample multiple posterior formulations $\pi(x^1, \theta^1), \ldots, \pi(x^K, \theta^K)$ for which the spatial input parameters $x^1, \ldots, x^K$ are coarsened to varying degrees. Each formulation runs its simulator $\eta^k(x^k)$ at its own particular grid resolution. This MCMC scheme, which borrows from the work of Goodman and Sokal (1989) and Liu and Sabatti (1999), allows information from the faster running, but less accurate, coarse formulations to speed up the mixing for the fine scale chains. In addition, this scheme is relatively easy to implement on a parallel environment, without having to "parallelize" the actual simulator code. This distributed, coupled MCMC approach is discussed in Section 3. Section 4 follows giving a final discussion.

2. SPATIAL REPRESENTATIONS

The simulator typically requires that x be input over m regular grid points at spatial locations denoted by the set $s^x = \{s_1^x, \ldots, s_m^x\}$, which is contained in the spatial domain $\mathcal{S}$. Hence the actual input to $\eta(\cdot)$ requires x be restricted to the grid points $x_{s^x} = (x(s_1^x), \ldots, x(s_m^x))^T$. As regards to notation, we use x when the process is only considered at the set of spatial locations s^x; we take $x(s)$ to mean that the process is defined for all $s \in \mathcal{S}$. The grid size m can often be specified in the simulator $\eta(x)$, with fine grids typically giving more accurate results at the cost of increased computation. We note that recent literature has stressed the importance of specifying spatial models that are consistent under coarsening or aggregation schemes. Clearly, the single component $x_{s_i^x}$ of the input grid x is some form of aggregate of a continuously defined process in the neighborhood of the location s_i^x. However, in many applications involving simulation of physical systems, aggregation – equivalently, upscaling or closure – is a difficult, or even an ill-posed task by itself. So, even though issues regarding aggreation consistency can play an important role, especially when the aggregation process is well defined and data are sufficiently informative, the data are usually insufficient to resolve x in much detail in the applications we consider. Hence we only require that the prior distribution for x infuse prior knowledge about its spatial distribution – at least at the resolution/level of detail that we expect from the data – as well as regularize the posterior for x.

In this paper we use both intrinsic Gaussian Markov random fields (MRFs) that model the m-dimensional process x at the spatial locations s^x, and standard Gaussian processes (GP) that define a process $x(s)$ over continuous space. The intrinsic Gaussian

MRF has the form

$$\pi(x|\theta_x) \propto \theta^{\frac{m}{2}} \exp\left\{-\tfrac{1}{2}\theta x^T W x\right\} \tag{1}$$

where θ_x controls the scale of x and the MRF precision matrix has the simple form

$$W_{ij} = \begin{cases} n_i \text{ if } i = j \\ -1 \text{ if } i \sim j \\ 0 \text{ otherwise} \end{cases} \tag{2}$$

where n_i is the number of neighbors of site s_i^x and $i \sim j$ means locations s_i^x and s_j^x are neighbors of one another. With the regular grids considered in this paper, we specify two sites s_i^x and s_j^x to be neighbors if they are directly adjacent on the grid so that interior points of a 2-d rectangular grid have 4 neighbors; edge sites have 3; and corner sites have 2.

Gaussian process priors are typically specified through their mean and covariance function. We take the mean to be constant and define the covariance by

$$\text{Cov}(x(s_i^x), x(s_j^x)) = \theta_1 \rho\left(\frac{\|s_i^x - s_j^x\|}{\theta_2}\right)$$

where the correlation function $\rho(\cdot)$ must be positive definite and satisfy $\rho(0) = 1$. We typically take $\rho(d) = e^{-d^2}$ which leads to very smooth realizations for $x(s)$. By contrast, realizations under the locally planar MRF model (2) exhibit local roughness.

This distinction is important if one wishes to infer about the local nature of x and if the data are informative about the small scale nature of x. It is often the case in inverse problems that the indirectly observed data give no information regarding the small-scale behavior of x. Also, the input grid x can best be regarded as the aggregate of an underlying continuous process. For the two reasons above it is often impossible to distinguish between locally smooth and locally rough character of $x(s)$ from the data alone. When this is the case, as it is in the hydrology examples, computational considerations can lead us to favor models with smooth local behavior.

When we can get away with a smooth GP specification for $x(s)$, we can then efficiently represent $x(s)$ by convolving a white noise process $u(s)$, $s \in \mathcal{S}$ with a smoothing kernel $k(s)$ so that

$$x(s) = \int_{\mathcal{S}} k(\nu - s)u(\nu)d\nu \text{ for } s \in \mathcal{S}. \tag{3}$$

The resulting covariance function for $x(s)$ depends on the displacement vector $d = s - s'$ and is given by

$$\text{Cov}(x(s), x(s')) \propto \rho(d) \propto \int_{\mathcal{S}} k(\nu - s)k(\nu - s')d\nu = \int_{\mathcal{S}} k(\nu - d)k(\nu)d\nu. \tag{4}$$

The proportionality depends on the scale of the white noise process $u(s)$ and on $\int_{\mathcal{S}} k^2(\nu)d\nu$. We typically take $\mathcal{S}$ to be $R^{1,2,\text{ or }3}$, and $k(\cdot)$ to be a normal density with independence between the coordinate component directions. This is an equivalent representation of a mean 0 GP with $\rho(d) = e^{-d^2}$, possibly after rescaling the coordinate axes. By restricting the latent process $u(s)$ to coarse lattice locations $s_1^u, \ldots, s_\ell^u$, a small number of parameters effectively control the entire process $x(s)$. Now with a discrete white noise process

$$u = (u(s_1^u, \ldots, u(s_\ell^u))^T \sim N(0, I_\ell/\theta_u)$$

$x(s)$ can be represented by the discrete analog of (3)

$$x(s) = \sum_{k=1}^{\ell} u_k k(s - s_k^u) \tag{5}$$

where $k(\cdot - s_k^u)$ is the smoothing kernel centered at s_k^u. Figure 1 shows three successively coarsened white noise realizations u, their induced processes $x(s)$ from convolving u with the kernel shown in the upper left of the top row of figures; the bottom row of figures shows $\mathrm{Cor}(x(s_0), x(s))$ as a function of s. The dotted black line gives the ideal covariance function obtained via (4). For this smooth process, u can undergo substantial coarsening before the induced process begins to substantially deviate from the ideal one obtained from continuous white noise.

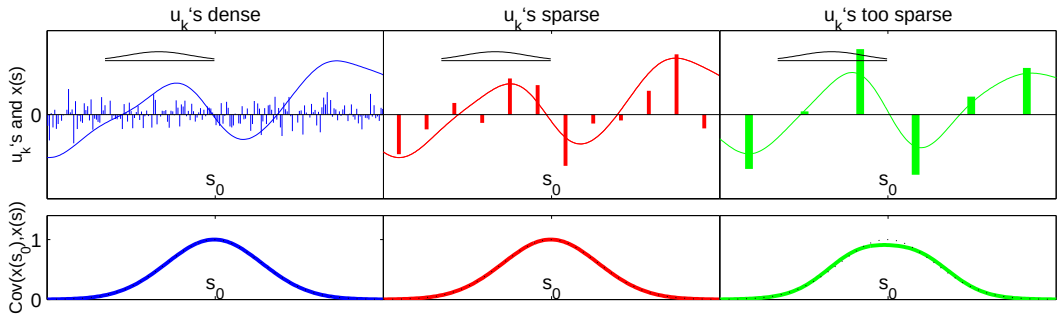

Figure 1. *A stationary spatial process $x(s)$ can be generated by smoothing white noise. The top frames the induced process $x(s)$ obtained by smoothing the white noise shown by vertical lines using the kernel shown in the top left of the figure. Moving from left to right, the underlying white noise process becomes successively coarser. Below each of the top frames is a function showing $Cov(x(s_0), x(s))$ as a function of s; the location of s_0 is marked in the figures. In the rightmost frame the u_k's are so sparse that the covariance of the induced process begins to deviate from the ideal covariance function it is trying to match, which is shown by the black dotted line.*

Before moving on, we note there are alternative lower dimensional representations of $x(s)$ that one may consider such as Cholesky, SVD, or Fourier. Taking x to be discrete, in each case we can express $x = Ku$ so that x is the weighted sum of bases given by the columns of K. The difference between the approaches is in the specification of K. We favor the moving average representation because of its local nature as well as the simplicity of its basis representation. Its local nature meshes well with MCMC in which a simple Metropolis update of individual u_k will influence a local region of $x(s)$. The simplicity of this basis representation easily allows for extending the basic model for which $u \sim N(0, I_\ell/\theta_u)$. By allowing more general dependence within u the model can be extended to account for non-stationarity or time dependence; see Calder *et al.* (2002) for example.

Example 1. Studying the flow of water underground is of great interest to engineers, with important applications to cleanup of contaminated soil and petroleum exploration and production. A statistically interesting component of this problem is the inverse problem of inferring soil structure (e.g., permeability) from flow data. Further details and references can be found in Lee *et al.* (2002).

The data presented here are from a larger study (Annable *et al.* 1998) at the Hill Air Force Base in Utah where the ground contains a number of contaminants. We

look only at conservative tracer data, an experiment that yields information only on the permeabilities and not on the contamination. The site is 14 feet by 11 feet, with four injection wells along one edge and three production (extraction) wells along the opposite edge. Water is pumped continuously through the field, and then a tracer is added and the time of travel is measured for the tracer from the injection wells to five measurements sites (sampling wells) in the field. This time of travel is referred to as the breakthrough time for each sampling location. Since water flows faster through regions of higher permeability, one can learn about the underlying permeabilities through the breakthrough times. The upper left plot of Figure 2 shows the locations of the injectors, producers, and samplers, with the breakthrough times shown for the sampling locations.

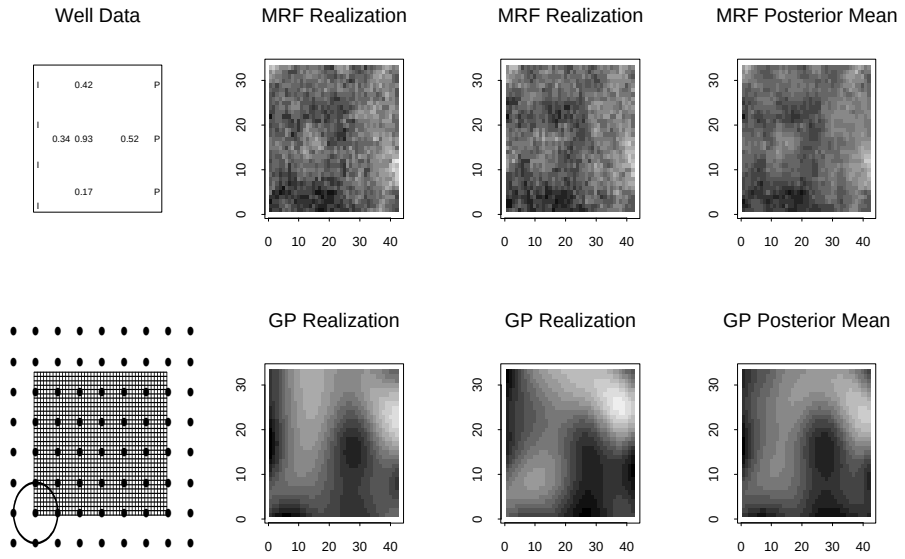

Figure 2. *Layout of wells, posterior realizations, and posterior means for both an MRF model and a moving average Gaussian Process model for the Hill Air Force Base data. In the upper left plot, the wells are labeled "I" for injectors, "P" for producers, and the samplers are shown with numbers where the value is the breakthrough time (in days) for each well. For the permeability plots, darker regions correspond to higher permeability values.*

Permeabilities vary spatially and are typically considered to be log-normally distributed. Thus all our priors for permeabilities are stated on the log scale. We use a 42 by 33 grid of square cells, one-third of a foot on each side. For notational convenience, we represent the unknown (log) permeabilities as a $m = 42 \times 33 \times 1$ lattice x. Conditional on a specified permeability field x, the breakthrough times are found from the solution of differential equations given by physical laws, i.e., conservation of mass, Darcy's Law, and Fick's Law. We do this using the S3D streamtube computer code of King and Datta-Gupta (1998) and find the $n = 5$ fitted breakthrough times, $\hat{y} = \eta(x)$.

We consider two formulations – one for which x is modeled as a 2-d MRF prior using four nearest neighbors on a $m = 42 \times 33$ lattice; and one for which x is parameterized as a GP via (5), where the $\ell = 72$ kernel locations are shown by the dots in the bottom left frame of Figure 2 and the kernels are bivariate normal with a one sd ellipse shown in the bottom left frame of Figure 2. The resulting posteriors are

$$\pi(x, \theta_x|y) \propto \exp\{-\tfrac{1}{2}\lambda(y-\eta(x))^T(y-\eta(x))\} \times \theta_x^{-\frac{m}{2}} \exp\{-\tfrac{1}{2}\theta_x x^T W x\} \times \theta^{\alpha x - 1} e^{-\beta x \theta}$$

$$\pi(u, \theta_u|y) \propto \exp\{-\tfrac{1}{2}\lambda(y-\eta(x))^T(y-\eta(x))\} \times \theta_u^{-\frac{\ell}{2}} \exp\{-\tfrac{1}{2}\theta_u u^T u\} \times \theta_u^{\alpha u - 1} e^{-\beta_u \theta_u}.$$

In the MCMC implementations, x is updated via multivariate Hastings steps (Lee *et al.* 2002) and u is updated via single site Metropolis steps.

Figure 2 shows the results from the two formulations. The top row shows two realizations from the posterior and the posterior mean for the MRF prior, while the bottom row shows analogous plots for the GP prior. Both models fit the observed data well. In particular, both show a region of lower permeability in front of (relative to the injectors) the central well, which has the latest breakthrough time. $\diamond$

3 COMPUTATION

3.1 *Linking coarse and fine formulations*

In many applications, the computational demands of the simulator greatly restrict the number of simulator runs that can be carried out, making posterior exploration via standard MCMC difficult or even impractical. An alternative is to formulate a coarsened version of the problem. Under this coarse specification, a coarsened counterpart for the input x is defined by $\tilde{x} = (\tilde{x}_1, \ldots, \tilde{x}_{\tilde{m}})^T = Cx$, where C is the coarsening operation which maps a m-vector to a lower dimensional $\tilde{m}$-vector. We use $s^{\tilde{x}} = \{s_1^{\tilde{x}}, \ldots, s_{\tilde{m}}^{\tilde{x}}\}$ to denote the spatial locations associated with this coarse grid. Typically, C is a $\tilde{m} \times m$ matrix so that Cx is a simple linear transformation, such as averaging or summing groups of fine-scale pixels to make coarse pixels. However, coarsening, or upscaling, could conceivably be a more complicated operation, depending on the application. Depending on the problem, y, θ_y, and θ_x might also require coarsened counterparts $\tilde{y}$, $\tilde{\theta}_{\tilde{y}}$, and $\tilde{\theta}_{\tilde{x}}$ which are modifications of their original form. In addition, the likelihood and priors under the coarsened formulation may also differ. The net result is two separate posterior distributions – one fine and one coarse:

$$\begin{aligned} \text{fine} \quad & \pi(x, \theta|y) \propto L(y|\eta(x), \theta) \times \pi(x|\theta_x) \times \pi(\theta) \\ \text{coarse} \quad & \tilde{\pi}(\tilde{x}, \tilde{\theta}|\tilde{y}) \propto \tilde{L}(\tilde{y}|\eta(\tilde{x}), \tilde{\theta}) \times \tilde{\pi}(\tilde{x}|\tilde{\theta}_{\tilde{x}}) \times \tilde{\pi}(\tilde{\theta}). \end{aligned}$$

In order to link the coarse and fine-scale formulations, we make use of *Metropolis coupled* MCMC (Geyer, 1991). Now, instead of running two separate MCMC chains, one on the fine posterior and one on the coarse posterior, a single chain is run on the product distribution. This coupled chain has stationary distribution $\pi(x, \theta|y) \times \tilde{\pi}(\tilde{x}, \tilde{\theta}|\tilde{y})$. Because of the coarsened input $\tilde{x}$ to the simulator, the chain sampling the coarse-scale posterior will run more quickly. In addition, the coarse-scale posterior is typically smoother and easier to sample via MCMC as compared to its fine-scale counterpart. Hence an efficient coupling scheme will allow information to move between the two formulations.

One possible implementation of such a coupled chain alternates standard within-scale updates with "swapping" updates that allow information to move between the two scales as shown below:

$$\begin{array}{ccccccccccc} (x,\theta)^1 & \overset{\text{MCMC}}{\longrightarrow} & (x,\theta)^2 & \overset{\text{SWAP}}{\longrightarrow} & (x,\theta)^3 & \overset{\text{MCMC}}{\longrightarrow} & (x,\theta)^4 & \overset{\text{SWAP}}{\longrightarrow} & (x,\theta)^5 & \cdots \\ (\tilde{x},\tilde{\theta})^1 & \overset{\text{MCMC}}{\longrightarrow} & (\tilde{x},\tilde{\theta})^2 & & (\tilde{x},\tilde{\theta})^3 & \overset{\text{MCMC}}{\longrightarrow} & (\tilde{x},\tilde{\theta})^4 & & (\tilde{x},\tilde{\theta})^5 & \cdots \end{array}$$

Here the updates denoted by $\overset{\text{MCMC}}{\longrightarrow}$ affect parameters within a given scale, while the updates denoted by $\overset{\text{SWAP}}{\longrightarrow}$ are a Hastings update that proposes new candidates $(x^*, \theta^*, \tilde{x}^*, \tilde{\theta}^*)$ according to the proposal kernel

$$q((x, \theta, \tilde{x}, \tilde{\theta}) \to (x^*, \theta^*, \tilde{x}^*, \tilde{\theta}^*)),$$

which is accepted according to the Hastings rule with probability

$$1 \wedge \frac{\pi(x^*, \theta^*|y)\tilde{\pi}(\tilde{x}^*, \tilde{\theta}^*|\tilde{y}) \times q((x^*, \theta^*, \tilde{x}^*, \tilde{\theta}^*) \rightarrow (x, \theta, \tilde{x}, \tilde{\theta}))}{\pi(x, \theta|y)\tilde{\pi}(\tilde{x}, \tilde{\theta}|\tilde{y}) \times q((x, \theta, \tilde{x}, \tilde{\theta}) \rightarrow (x^*, \theta^*, \tilde{x}^*, \tilde{\theta}^*))} \tag{6}$$

where $a \wedge b$ is the minimum of a and b.

We now describe some specific swapping proposals $q((x, \theta, \tilde{x}, \tilde{\theta}) \rightarrow (x^*, \theta^*, \tilde{x}^*, \tilde{\theta}^*))$ for the applications we consider. It is often convenient to break the swapping proposal kernel into the product

$$q((x, \theta, \tilde{x}, \tilde{\theta}) = q((x, \theta) \rightarrow (\tilde{x}^*, \tilde{\theta}^*)) \times q((\tilde{x}, \tilde{\theta}) \rightarrow (x^*, \theta^*))$$

where $q((x, \theta) \rightarrow (\tilde{x}^*, \tilde{\theta}^*))$ generates a coarse-scale proposal $(\tilde{x}^*, \tilde{\theta}^*)$ from the current fine-scale state (x, θ), and the kernel $q((\tilde{x}, \tilde{\theta}) \rightarrow (x^*, \theta^*))$ generates a fine-scale proposal (x^*, θ^*) from the current coarse-scale state $(\tilde{x}, \tilde{\theta})$.

Swapping proposals for MRF priors. When we use the MRF prior for x and $\tilde{x}$ (1), we generate the coarse-scale proposal by deterministically coarsening the fine-scale state and then generating a candidate value $\tilde{\theta}^*$ by simulating from its full conditional distribution (under the coarse-scale posterior) given the new proposed value $\tilde{x}^*$. This proposal kernel can be written

$$q((x, \theta) \rightarrow (\tilde{x}^*, \tilde{\theta}^*)) = I[\tilde{x}^* = Cx] \times \tilde{\pi}(\tilde{\theta}^*|\tilde{x}^*, \tilde{y})$$

where $I[\cdot]$ is the indicator function, Cx is the coarsening operation applied to the fine-scale x, and $\tilde{\pi}(\tilde{\theta}|\tilde{x}^*, \tilde{y})$ is the full conditional distribution of $\tilde{\theta}$ under the coarse formulation. If $\tilde{\theta}$ is given a conjugate $\Gamma(\alpha_x, \beta_x)$ prior, then its full conditional also has a gamma form. Also for the applications we consider, C is a simple summing or averaging operation.

The fine-scale candidate (x^*, θ^*) given the current coarse-scale state $(\tilde{x}, \tilde{\theta})$ is generated by drawing from the prior distribution $\pi(x^*|\theta^\dagger)$ subject to the constraint $Cx^* = \tilde{x}$. The value $\theta^\dagger$ is a deterministic function of $\tilde{\theta}$ chosen so that the candidate x^* most nearly matches the properties of typical fine-scale realizations. In the 1-d application of Example 2, we take $\theta^\dagger = \frac{1}{8}\tilde{\theta}$; in the 2-d SPECT application of Example 3, we take $\theta^\dagger = \frac{1}{32}\tilde{\theta}$. Once x^* has been generated, θ^* can then be drawn from its full conditional given the candidate value x^*. Hence

$$q((\tilde{x}, \tilde{\theta}) \rightarrow (x^*, \theta^*)) \propto \pi(x^*|\theta^\dagger, y)I[\tilde{x} = Cx^*] \times \pi(\theta^*|x^*, y).$$

Note that when the prior $\pi(x|\theta)$ has a multivariate normal form and C is a matrix, then the proposal x^* can be generated directly. This update is more problematic when the prior for x is not normal.

Example 2 Before considering swapping updates for formulations involving moving average specifications for x, we first consider a synthetic blur free, 1-d imaging example. A smooth, 1-d souce is emitting according to a Poisson process with intensity given by the smooth, solid line(s) in Figures 3 (a & b). Under the fine-scale formulation of the problem is a 1-d array of $n = 40$ detectors recording emissions from the source; the count for each detector is shown in Figure 1(a). A 1-d Gaussian MRF prior over the $m = 40$ detector sites is assigned to the unknown fine-scale image x, with a $\Gamma(\alpha, \beta)$ prior for the precision parameter θ_x. A coarse-scale formulation is obtained by combining

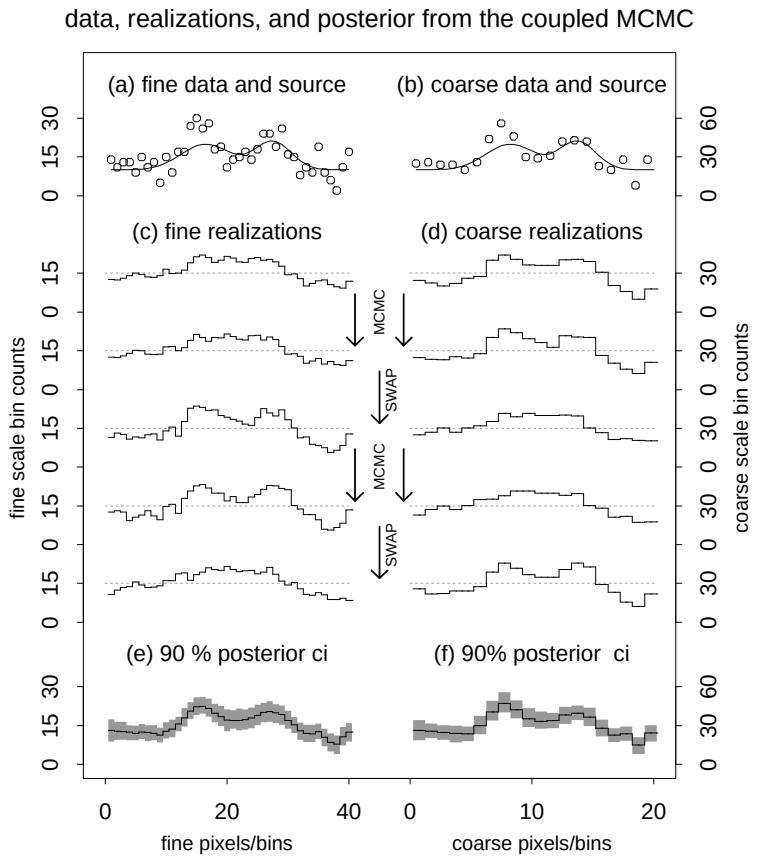

Figure 3. *Data, posterior realizations, and posterior summary for the coupled MCMC scheme: (a & b) Data and true image intensity under the fine and coarse formulations; (c & d) a sequence of four updates under the coupled MCMC scheme; (e & f) pointwise posterior 90% credible intervals for the image intensities x and $\tilde{x}$ under the fine and coarse formulations.*

adjacent detector pairs so that the coarsened data consist of $\tilde{n} = 20$ counts (Figure 1(b)). Similarly, a MRF prior is assigned to the coarsened image $\tilde{x}$, which is divided into $\tilde{m} = 20$ sites, one for each coarse detector.

The fine and coarse formulations are given by

$$\begin{array}{cc} \text{fine} & \text{coarse} \\ L(y|x) \propto \prod_{i=1}^{n} x_i^{y_i} \exp\{-x_i\} & \tilde{L}(\tilde{y}|\tilde{x}) \propto \prod_{i=1}^{\tilde{n}} \tilde{x}_i^{\tilde{y}_i} \exp\{-\tilde{x}_i\} \\ \pi(x|\theta) \propto \theta^{\frac{m}{2}} \exp\{-\frac{1}{2}\theta x^T W x\} & \tilde{\pi}(\tilde{x}|\tilde{\theta}) \propto \tilde{\theta}^{\frac{\tilde{m}}{2}} \exp\{-\frac{1}{2}\tilde{\theta}\tilde{x}^T \tilde{W} \tilde{x}\} \\ \pi(\theta) \propto \theta^{\alpha-1} e^{-\beta\theta} & \tilde{\pi}(\tilde{\theta}) \propto \tilde{\theta}^{\alpha-1} e^{-\beta\tilde{\theta}} \end{array}$$

where W and $\tilde{W}$ are given by (2) and adjacent detectors are defined to be neighbors. The swapping updates are carried out as described previously, with $\theta^{\dagger} = \frac{1}{8}\tilde{\theta}$. Figures 3 (c & d) show four successive updates from this coupled MCMC scheme. The resulting posterior pointwise 90% credible intervals are shown in Figures 3 (e & f) under both the fine and coarse formulations. In this application, the swap proposals were accepted about 14% of the time. $\diamond$

Example 3. In SPECT the goal is to estimate a photon emission intensity map using photon emissions from an object detected by a gamma camera. Figure 4 diagrams the

information obtained during a SPECT scan. As the object emits photons, the gamma camera records the locations of photon hits along the camera array. The gamma camera array can rotate completely around the object. At a given camera position, photon emissions are recorded as counts at each of 128 bins indexed by b. This accumulation of counts is repeated at each of 120 rotation angles indexed by angle a.

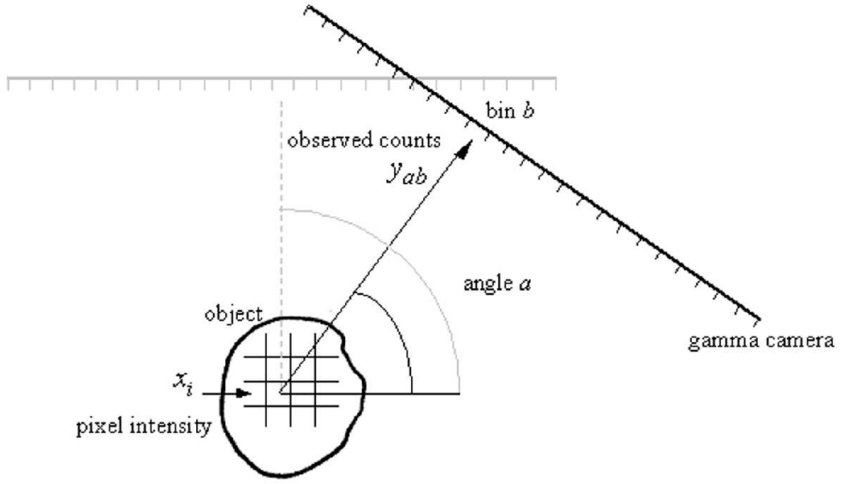

Figure 4. *SPECT: An object emits photons with location dependent intensity $x(s)$. The gamma camera obtains binned counts of photon emissions from various different positions controlled by the angle a. The counts from each angle a and each bin of the gamma camera b are recorded as y_{ab}.*

The data consist of counts y_{ab} obtained from bin b of the gamma camera while it was positioned at angle a. Lead columnators on the camera ensure that photons hit the camera at nearly right angles. Since a photon may be scattered, absorbed, miss the gamma camera, or otherwise fail to be detected, the probability map p_{abi} gives the probability of an emission from pixel i being detected at angle a and bin b.

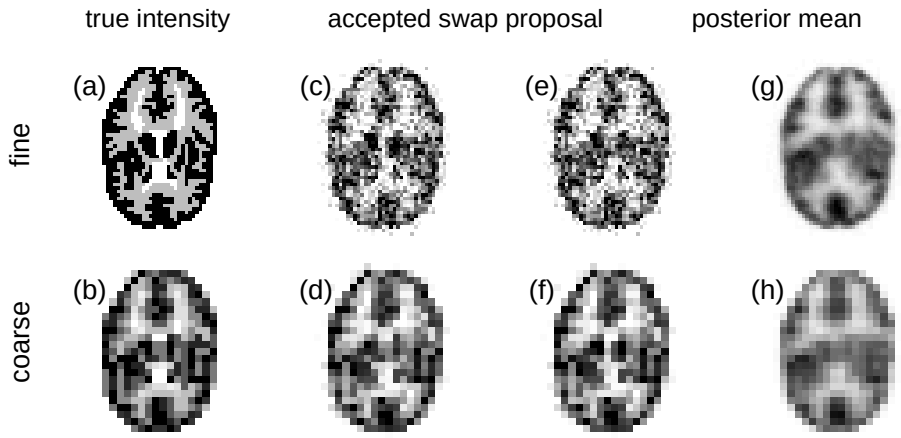

Figure 5. *Coupled fine and coarse-scale MCMC for a SPECT example. (a) true emission intensities; (b) coarsened version of the true intensities; (c & d) current values for x and $\tilde{x}$ during the coupled MCMC run; (e & f) proposed fine and coarse images x^* and $\tilde{x}^*$ after swapping an interior patch of the images in (c & d); (g) posterior mean for x; (h) posterior mean for $\tilde{x}$.*

We specify a fine-scale formulation that divides the emission source into a $m = 128 \times 128$ lattice of pixels; the coarse-scale formulation divides the emission source into a $\tilde{m} = 64 \times 64$ lattice of pixels. Hence the fine formulation requires a $120 \times 128 \times 128^2$ probability map p_{abi}, and the coarse formulation requires a $120 \times 128 \times 64^2$ probability map $\tilde{p}_{abi}$, The counts y_{ab} then have a Poisson distribution with mean λ_{ab} under the fine-scale formulation, and mean $\tilde{\lambda}_{ab}$ under the coarse-scale formulation where

$$\lambda_{ab} = \sum_{i=1}^{m} x_i p_{abi} \text{ and } \tilde{\lambda}_{ab} = \sum_{i=1}^{\tilde{m}} \tilde{x}_i \tilde{p}_{abi}.$$

Hence computing changes in λ_{ab} due to changing a component of x requires four times as much effort as does computing changes in $\tilde{\lambda}_{ab}$ due to changing a component of $\tilde{x}$.

The two formulations can then be written

fine	coarse
$L(y\|x) \propto \prod_{a,b} \lambda_{ab}^{y_{ab}} \exp\{-\lambda_{ab}\}$	$\tilde{L}(y\|\tilde{x}) \propto \prod_{a,b} \tilde{\lambda}_{ab}^{y_{ab}} \exp\{-\tilde{\lambda}_{ab}\}$
$\pi(x\|\theta) \propto \theta^{\frac{m}{2}} \exp\{-\frac{1}{2}\theta x^T W x\}$	$\tilde{\pi}(\tilde{x}\|\tilde{\theta}) \propto \tilde{\theta}^{\frac{\tilde{m}}{2}} \exp\{-\frac{1}{2}\tilde{\theta}\tilde{x}^T \tilde{W} \tilde{x}\}$
$\pi(\theta) \propto \theta^{\alpha-1} e^{-\beta\theta}$	$\tilde{\pi}(\tilde{\theta}) \propto \tilde{\theta}^{\alpha-1} e^{-\beta\tilde{\theta}}$

where W and $\tilde{W}$ are given by (2) with vertically and horizontally adjacent pixels defined as neighbors. Note that the data are not coarsened in this example. Within-scale MCMC is carried out as described in Weir (1997).

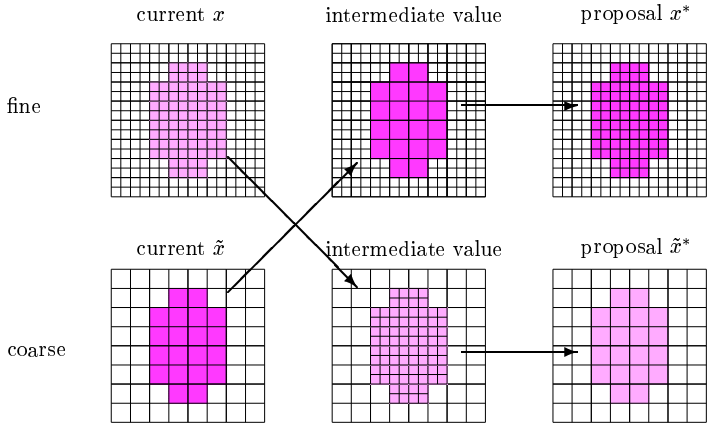

Figure 6. *A proposal that swaps only a piece of the image between the coarse- and fine-scales. Given the current values for x and $\tilde{x}$, the shaded regions of the two images are exchanged giving the intermediate values. The coarse shaded piece is refined to give a fine proposal x^* and the fine shaded piece is coarsened to give a coarse proposal $\tilde{x}^*$. The stochastic refining of the coarse shaded piece conditions on its previous coarse value as well as its neighboring fine-scale pixels.*

We originally used swaps as described in Section 3.1. However we found that the fine-scale proposals were not sufficiently accurate near the edges of the emission phantom (Figure 5 (a)). Instead we proposed to swap only interior pieces of the fine and coarse images. Figure 6 shows how this is carried out. To construct the proposal, the same interior regions of the two images are exchanged. The region within the

coarse-scale exterior is then deterministically coarsened; the region within the fine-scale exterior is then refined, conditioned on matching its coarse values and conditioned on the fine-scale pixels neighboring the region. This gives the proposal a better chance of being accepted – about one in eight swap proposals are accepted. Figure 5 shows an accepted swap along with coarse and fine-scale posterior mean images.

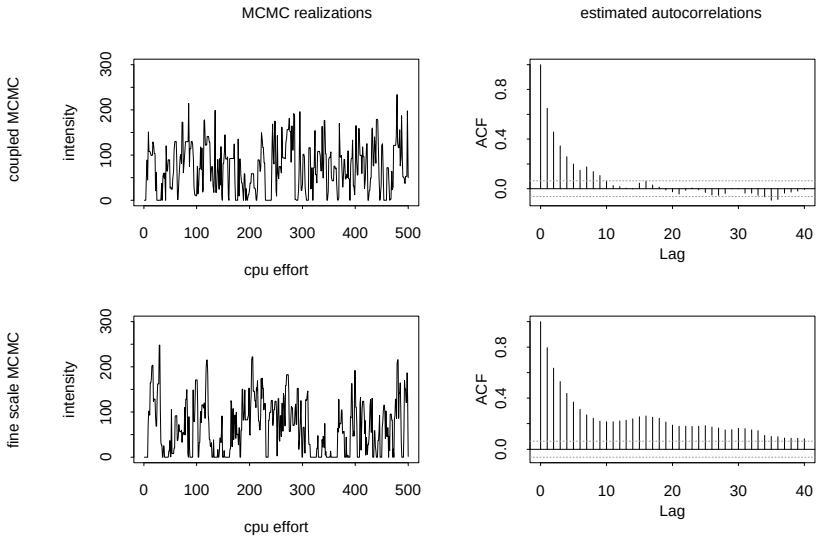

Figure 7. *MCMC trace plots and autocorrelation plots for the intensity of an interior pixel in the SPECT application under the coupled MCMC approach (top row) and standard MCMC within the fine-scale only (bottom row). The trace plots are standardized to comparable CPU time. The coupled MCMC is about three times as efficient when standardized to CPU effort.*

MCMC trace plots are shown in Figure 7 for an interior pixel under the two posterior sampling schemes. The coupled MCMC yields estimated autocorrelation times that are about a third of those obtained under the standard fine-scale MCMC algorithm. ⋄

Swapping proposals for continuous spatial priors.

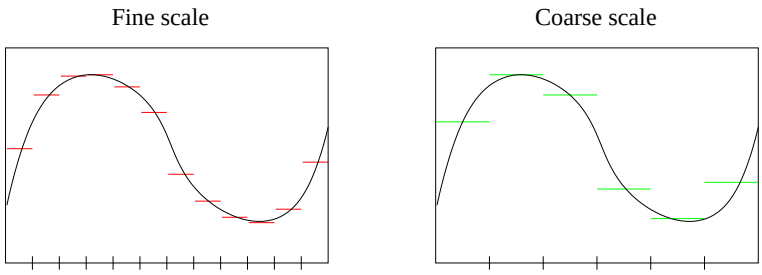

Figure 8. *Deterministic coarsening and refining in the case when x and $\tilde{x}$ are modeled as restrictions of identically distributed continuous processes $x(s)$ and $\tilde{x}(s)$. Given a realization of the underlying continuous process, the restriction of the process to the fine locations s^x or the coarse locations $s^{\tilde{x}}$ is completely determined.*

In the case when x and $\tilde{x}$ are both modeled a priori as restrictions of an identically distributed continuous processes $x(s)$ and $\tilde{x}(s)$ the swapping is trivial. These processes are constructed via (5) using independent copies u and $\tilde{u}$ with common spatial locations

$s^u = \{s_1^u, \ldots, s_\ell^u\}$ so that

$$x(s) = \sum_{k=1}^{\ell} u_k k(s - s_k^u) \quad \text{and} \quad \tilde{x}(s) = \sum_{k=1}^{\ell} \tilde{u}_k k(s - s_k^u)$$

with u and $\tilde{u}$ modeled as independent $N(0, I_\ell/\theta)$ and $N(0, I_\ell/\tilde{\theta})$ draws, respectively. The gridded x essentially represents the continuous process as a piecewise constant over pixels centered at the locations s^x. Likewise $\tilde{x}$ represents $x(s)$ as piecewise constants over larger pixels centered at the coarse locations $s^{\tilde{x}}$ (Figure 8).

A swap between x and $\tilde{x}$ can be carried out by simply exchanging the values of (u, θ) and $(\tilde{u}, \tilde{\theta})$. Hence, coarsening x amounts to evaluating $x(s)$ at the coarse locations $s^{\tilde{x}}$; refining $\tilde{x}$ amounts to evaluating $\tilde{x}(s)$ at the fine locations s^x. Since this swap transition is symmetric and deterministic, the acceptance probability of (6) simplifies to a Metropolis acceptance rule. We defer to Section 3.2 to show an example of swapping using the continuous formulation for multiple levels of coarsening.

3.2 *Multi-processor implementation*

Perhaps the most appealing aspect of this coupled MCMC approach is that it is readily amenable to multiprocessor implementation, without having to "parallelize" the simulator code. Multiprocessor implementation is most easily carried out by running separate chains on the various processors, each exploring its own, possibly coarsened, posterior formulation. These chains are then coupled by periodically proposing swaps between the parameter values of the various chains as described in the previous section.

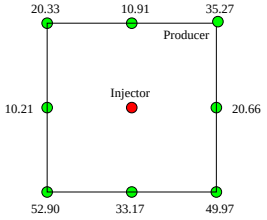

Figure 9. *Data for the multiprocessor sampler shown in Figure 10. An inverted nine spot pattern of a single injection well surrounded by 8 production wells. A shock of tracer is introduced at the injection well and the tracer concentration is recorded as a function of time at the production wells. The tracer breakthrough times are shown for each of the production wells. The likelihood is based on the breakthrough times.*

As an example we consider a synthetic application similar to the 2-d application of Section 2 where wells are laid out in an inverted nine spot pattern with a single injection well in the center surrounded by eight production wells. After a shock of tracer is introduced at the central production well, tracer breakthrough times are recorded at the eight production wells (Figure 9). Figure 10 shows an example of a three processor implementation with each processor running its own chain – one sampling a coarse-scale posterior, one sampling an intermediate-scale posterior, and one sampling a fine-scale posterior. The multiprocessor sampler alternates between within-scale updates and swapping updates. The within-scale updates consist of four MCMC scans of the permeability image at the coarse-scale, 2 scans at the intermediate-scale, and one scan at the fine level. The swapping scans consist of proposing swaps between current permeability images for each of the three possible scale pairings (coarse-intermediate, coarse-fine, and intermediate-fine). An implementation involving 7 differnt levels of resolution was also carried out where swaps were attempted between all levels of coarseness. The proportion of accepted swaps are summarized in Table 1.

Though this example demonstrates a practical parallel implementation of a multigrid MCMC scheme, clearly a number of questions loom regarding: allocation of processors to formulations; the choice of levels of coarseness in the auxiliary formulations;

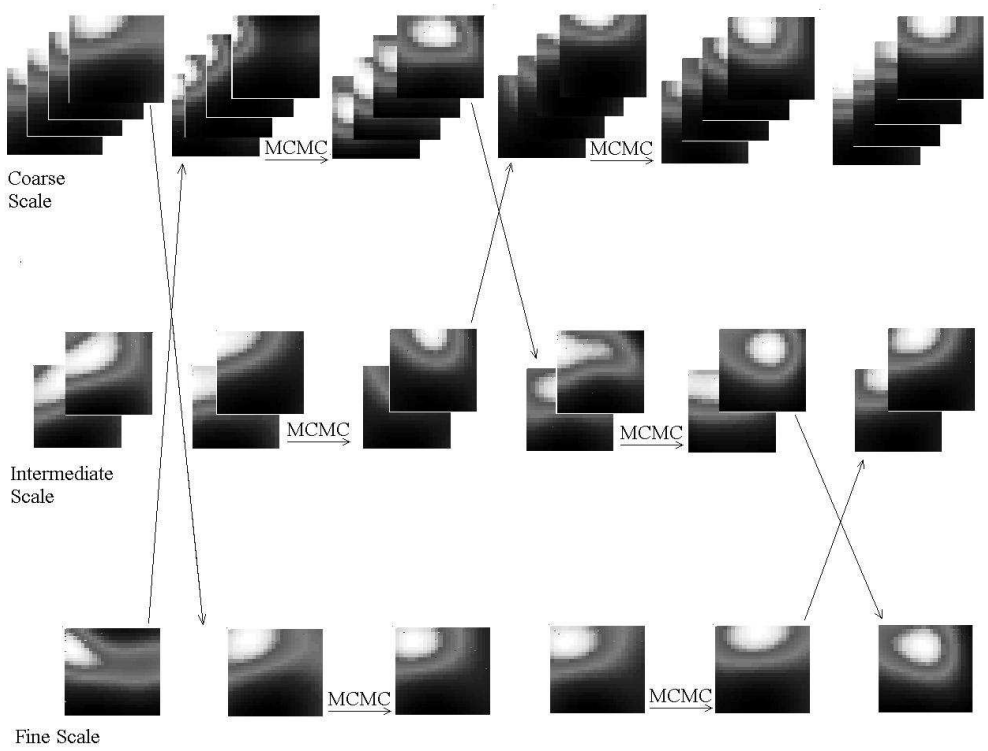

Figure 10. *Running formulations at different scales on different processors. Three distinct posterior distributions are obtained for the hydrology application of Section 2.2 by using different grid sizes in the flow simulator (*16×16, 20×20, *and* 24×24*). The three resulting posteriors are then sampled on three distinct processors. After a within-scale update scan consisting of 4 MCMC scans on the coarse-scale formulation, 2 on the intermediate-scale, and 1 on the fine-scale, metropolis swaps are proposed between the current realizations at each processor. This figure shows realizations at each level of coarseness for successive within-scale updates along with the result of the metropolis swaps between scales. Three such sequences are shown. In the first, the arrows denote an accepted swap between the coarse and fine-scales, in the second, a coarse-intermediate swap is accepted, in the third, an intermediate-fine swap is accepted. The* $\overset{\text{MCMC}}{\longrightarrow}$ *symbol denotes 3 additional within-scale update scans.*

Table 1. *acceptance rates of swap proposals*

	28×28	24×24	20×20	16×16	12×12	8×8
32×32	0.86	0.70	0.39	0.13	0.01	0
28×28	-	0.80	0.47	0.22	0.01	0
24×24	-	-	0.69	0.30	0.03	0
20×20	-	-	-	0.56	0.11	0
16×16	-	-	-	-	0.21	0
12×12	-	-	-	-	-	0.01

and appropriate swapping strategies, just to name a few. We have found the guidelines in Geyer and Thompson (1995) and Liu and Sabatti (1999) regarding constructing augmented chains relevant here. Future work and additional experience will give us a better handle on such questions.

4. DISCUSSION

Distributed computing, stingy parameterization, and augmentation with additional fast, coarsened formulations has expanded the universe of inverse/model calibration problems that can be handled using MCMC for posterior exploration. This is particularly relevant since distributed machines, such as relatively cheap clusters of workstations, are becoming more common and more accessable. Implementation of the MCMC schemes proposed here are straightforward and require minimal knowledge in programming for distributed architectures.

We note that the use of Geyer's coupled MCMC could be replaced with simulated tempering as in Geyer and Thompson, using reversible jump MCMC (Green 1995) to handle the change in dimension that comes in moving between scales. Our use of coupled MCMC allows us to control the amount of processing on each scale and makes it unnecessary to compute normalizing constants.

REFERENCES

Andersen, K. E., Brooks, S. P. and Hansen, M. B. (2001). Bayesian inversion of geoelectrical resistivity data. *Tech. Rep.*, Dept. Math. Sci., Aalborg Univ..

Annable, M. D., Rao, P. S. C., Hatfield, K., Graham, W. D., Wood, A. L. and Enfield, C. G. (1998). Partitioning tracers for measuring residual napl: field-scale test results. *J. Env. Eng.* **124**, 498–503.

Calder, C., Holloman, C. and Higdon, D. (2003). A space-time model for ozone concentration using process convolutions. , (to appear). New York: Springer.

Craig, P. S., Goldstein, M., Rougier, J. C. and Seheult, A. H. (2001). Bayesian forecasting using large computer models. *J. Amer. Statist. Assoc.* **96**, 717–729.

Geyer, C. J. (1991). Monte carlo maximum likelihood for dependent data. *in* E. Keramidas (ed.), *Comp. Sci. and Statis.: Proc. 23rd Symp. Interface*, 156–163.

Geyer, C. J. and Thompson, E. A. (1995). Annealing markov chain Monte Carlo with applications to ancestral inference. *J. Amer. Statist. Assoc.* **90**, 909–920.

Goodman, J. and Sokal, A. D. (1989). Multigrid Monte Carlo method. *Phys. Rev. Let. D* **40**, 2035–2072.

Green, P. J. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika* **82**, 711–732.

Hegstad, B. K. and Omre, H. (2001). Uncertainty in production forecasts based on well observations, seismic data and production history. *Soc. Petrol. Eng. J.* 409–424.

Higdon, D., Lee, H. and Bi, Z. (2002). A Bayesian approach to characterizing uncertainty in inverse problems using coarse and fine scale information, *IEEE Trans. in Sig. Proc.*, (to appear). .

Kennedy, M. and O'Hagan, A. (2001). Bayesian calibration of computer models (with discussion). *J. Roy. Statist. Soc. B* **68**, 425–464.

King, M. and Datta-Gupta, A. (1998). Streamline simulation: A current perspective, *In Situ* **22**, 91–140.

Lee, H., Higdon, D. M., Bi, Z., Ferriera, M. and West, M. (2002). Markov random field models for high-dimensional parameters in simulations of fluid flow in porous media. *Technometrics*, (to appear).

Liu, J. and Sabatti, C. (1999). Simulated sintering: Markov chain Monte Carlo with spaces of varying dimensions (with discussion). *Bayesian Statistics 6* (J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith, eds.). Oxford: University Press, 386–413.

Oliver, D., Cunha, L. and Reynolds, A. (1997). Markov chain Monte Carlo methods for conditioning a permeability field to pressure data. *Math. Geol.* **29**, 61–91.

Sacks, J., Welch, W. J., Mitchell, T. J. and Wynn, H. P. (1989). Design and analysis of computer experiments (with discussion). *Statist. Sci.* **4**,409–423.

Weir, I. (1997). Fully Bayesian reconstructions from single photon emission computed tomography. *J. Amer. Statist. Assoc.* **92**, 49–60.